(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平10－283365

(43)公開日　平成10年(1998)10月23日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/401	３２０Ｚ
17/21			15/20	５９０Ｅ
			15/40	３７０Ａ
			15/403	３４０Ｂ
				３８０Ｄ

審査請求　未請求　請求項の数11　ＯＬ　（全 12 頁）

(21)出願番号　特願平9－90385

(22)出願日　平成９年(1997)４月９日

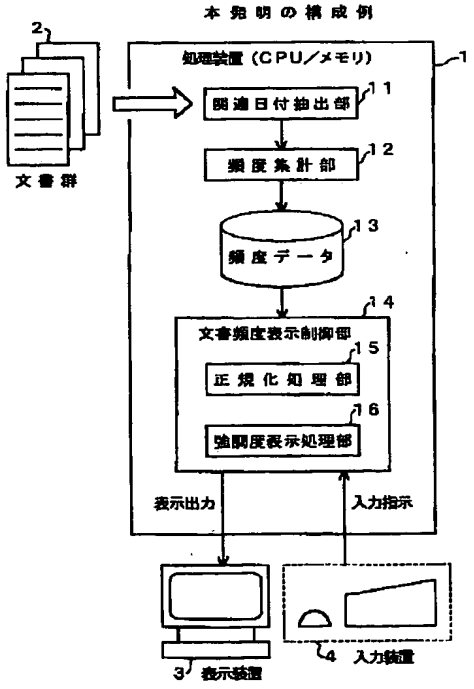
(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号

(72)発明者　津田　宏
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(74)代理人　弁理士　小笠原　吉義　（外２名）

(54)【発明の名称】　文書表示装置およびプログラム記憶媒体

(57)【要約】
【課題】非定型の文書群を整理して提示する文書表示装置に関し，文書内容に記述された日付情報を用いて，複数の時間的観点から，文書頻度を強調して提示することにより，多量の文書を容易に把握できるようにする。
【解決手段】関連日付抽出部11は，文書内容に記述されている日付表現の文字列をパターンマッチにより抽出し，日付情報を得る。頻度集計部12は，日付情報ごとの関連する文書数を集計して頻度データ13とする。文書頻度表示制御部14は，頻度データ13をもとに，正規化処理部15によって指定された時間単位ごとの文書数の分布を正規化して強調度を計算し，強調度表示処理部16によって，年／月／曜日／日といった時間単位ごとに強調度に応じた文書頻度の表示を行う。



【特許請求の範囲】
【請求項１】　電子化された文書群の情報を提示する文書表示装置であって，前記電子化された文書群の各文書に関連する日付情報を抽出する関連日付抽出手段と，抽出した日付情報に基づき，ある時間単位ごとに関連する文書の数を集計する頻度集計手段と，集計結果に基づき，ある時間単位ごとの文書の頻度情報を表示する文書頻度表示制御手段とを備えることを特徴とする文書表示装置。
【請求項２】　請求項１記載の文書表示装置において，前記関連日付抽出手段は，文書内容に記載された日付情報をパターンマッチにより自動的に抽出することを特徴とする文書表示装置。
【請求項３】　請求項１記載の文書表示装置において，前記文書の頻度情報を表示する時間単位を指定する入力手段を備え，前記頻度集計手段は，年，月，曜日または日といった時間単位ごとに関連する文書の数を集計し，前記文書頻度表示制御手段は，前記入力手段によって指定された年，月，曜日または日といった時間単位ごとの文書の頻度情報を表示することを特徴とする文書表示装置。
【請求項４】　請求項１記載の文書表示装置において，前記文書頻度表示制御手段は，前記時間単位における文書頻度の分布に基づき，各文書頻度を正規化して強調度に変換し，強調度に応じた表示を行うことを特徴とする文書表示装置。
【請求項５】　請求項１記載の文書表示装置において，前記文書頻度表示制御手段は，利用者の指定によりまたは自動的に文書頻度を強調度に変換する変換関数を切り換えて強調度を算出することを特徴とする文書表示装置。
【請求項６】　請求項４または請求項５記載の文書表示装置において，前記文書頻度表示制御手段は，算出した強調度に応じて，文書頻度を表す記号または図形の色，明度または大きさを変えて頻度情報を表示することを特徴とする文書表示装置。
【請求項７】　文書に関するデータベースの検索機能を有するシステムにおいて，データベースの検索結果としての文書群から，それらの各文書に関連する日付情報を抽出する関連日付抽出手段と，抽出した日付情報に基づき，ある時間単位ごとに関連する文書の数を集計する頻度集計手段と，集計結果に基づき，ある時間単位ごとの文書の頻度情報を表示する文書頻度表示制御手段とを備えることを特徴とする文書表示装置。
【請求項８】　共有する文書に識別子を付与して蓄積する手段と，共有文書に対するインデックスを登録する手段と，登録したインデックスに基づいて共有文書に関する情報を可視化するビュー生成手段とを有し，ネットワークを利用して文書を共有し整理するシステムにおいて，前記インデックス登録手段は，共有文書群の各文書に関連する日付情報を抽出する関連日付抽出手段と，抽出した日付情報に基づき，ある時間単位ごとに関連する文書の数を集計する頻度集計手段とを持ち，集計結果の頻度データをインデックスとして登録し，前記ビュー生成手段は，インデックスとして登録された頻度データに基づき，ある時間単位ごとの文書の頻度情報を表示する文書頻度表示制御手段を持つことを特徴とする文書表示装置。
【請求項９】　電子化された文書群の情報を提示する文書表示装置を実現するプログラムを記憶した媒体であって，前記電子化された文書群の各文書に関連する日付情報を抽出する関連日付抽出手段と，抽出した日付情報に基づき，ある時間単位ごとに関連する文書の数を集計する頻度集計手段と，集計結果に基づき，ある時間単位ごとの文書の頻度情報を表示する文書頻度表示制御手段とを実現するプログラムを格納したプログラム記憶媒体。
【請求項１０】　文書に関するデータベースの検索機能を有するシステムにおける文書表示装置を実現するプログラムを記憶した媒体であって，データベースの検索結果としての文書群から，それらの各文書に関連する日付情報を抽出する関連日付抽出手段と，抽出した日付情報に基づき，ある時間単位ごとに関連する文書の数を集計する頻度集計手段と，集計結果に基づき，ある時間単位ごとの文書の頻度情報を表示する文書頻度表示制御手段とを実現するプログラムを格納したプログラム記憶媒体。
【請求項１１】　共有する文書に識別子を付与して蓄積する手段と，共有文書に対するインデックスを登録する手段と，登録したインデックスに基づいて共有文書に関する情報を可視化するビュー生成手段とを有し，ネットワークを利用して文書を共有し整理するシステムにおける文書表示装置を実現するプログラムを記憶した媒体であって，前記インデックス登録手段の処理として，共有文書群の各文書に関連する日付情報を抽出する処理と，抽出した日付情報に基づき，ある時間単位ごとに関連する文書の数を集計する処理とを行い，集計結果の頻度データをインデックスとして登録する処理と，前記ビュー生成手段の処理として，インデックスとして登録された頻度データに基づき，ある時間単位ごとの文書の頻度情報を表示する処理とを実現するプログラムを格納したプログラム記憶媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は，非定型の文書群を整理して提示する文書表示装置，特に各日付において日付に関連する文書が出現する頻度を，年／月／曜日／日といった複数の時間的観点から整理して提示できるようにした文書表示装置およびその文書表示装置を実現するプログラム記憶媒体に関する。
【０００２】昨今のパーソナルコンピュータおよびネッ

トワークの発達により，テキストファイル，電子メールまたはネットワークを用いて配信されるニュースデータ（ネットワークニュース）など，個人が授受する電子化文書は増加する一途である。このような文書群を，その文書内容に基づいて整理して提示する技術が求められている。

【０００３】これらの文書群の内容には，例えば講演会の案内の日時，原稿の締切りなど，文書に関連する日付や時刻（日付情報）が記述されているものが多い。

【０００４】

【従来の技術】本発明は，広くはデータの可視化技術に関連する。従来，データの可視化は，データベースなどの定型情報に対して行われることが多かった。定型情報であれば，特定のフィールドを取り出して，既存のグラフ化ソフトウェア等と組み合わせることにより，所望のデータを可視化することができる。

【０００５】しかし，テキストファイル，電子メールやネットワークニュースなどの文書は，必ずしも特定の形式で情報が格納されているとは限らない。オペレーティング・システムにより，ファイルの属性という形で，ファイル名，ファイルサイズ，作成日付，作成者などの情報は文書に付加されて提供されるが，これだけでは文書の内容を表しているとはいえない。例えば，講演会の案内についての文書があったとしても，講演会がいつ開催されるのかは，実際に文書を読むことでしか知ることができない。

【０００６】また，インターネットなどを通じた非定型文書情報の爆発的な増加により，例えばインターネットＵＲＬの検索エンジンのように，文字列レベルで文書を検索する全文検索の技術が重要となっている。しかし，ただ単に，ある文字列が記述されているからといって，必ずしもその文字列が存在する文書が，求める文書であるとは限らない場合があり，検索結果にはノイズが含まれることになる。

【０００７】さらに，文書を検索する場合にも，検索対象となる文書の全体量が増大していること，加えて，検索漏れを少なくするために類似語を論理和（ＯＲ）で展開して検索することなどが要因で，検索した結果の文書量も膨大になってきている。このように，膨大な量でノイズの多い検索結果の中から有用な情報を探す上でも，データの可視化の技術は重要になっている。

【０００８】このような状況下において，従来，種々の文書群から文書の内容に基づいて，ある出来事がいつごろ多く発生しているか，ある事象の発生が年／月／曜日／日によってどのように変動するかなどを，わかりやすく可視化する装置はなかった。

【０００９】

【発明が解決しようとする課題】以上のように，従来，種々の文書群の中から特定の事柄についての検索技術はあったが，文書の時間的頻度情報を可視化する装置はなかった。膨大な数の文書情報の中から，特定の事柄に着目して，例えば食中毒は何月に多く発生する傾向があるか，交通事故の発生する割合は，曜日によってどのように変わるかなどを知ることができれば，統計的に有用な情報を得ることができると考えられる。

【００１０】本発明は，以上の点に鑑み，次のような従来の問題を解決する。１．従来技術では，非定型な文書は，ファイルの属性（ファイル名，ファイルサイズ，更新日）によってしか把握できず，また，内容により把握するためには，文書内や文書に関して特定の形式で情報を付加する必要があった。

【００１１】２．非定型の文書の関連する日付情報は，文書内容を実際に読むことでしか得られなかった。３．文書に関連する日付情報のある期間ごとの分布を，容易に把握することができなかった。

【００１２】本発明は上記問題点の解決を図り，電子化された文書群における日付に関連する文書の頻度を，年／月／曜日／日といった複数の時間的観点から整理して，わかり易く提示する手段を提供することを目的とする。

【００１３】

【課題を解決するための手段】図１は，本発明の構成例を示す図である。１はＣＰＵおよびメモリ等からなる処理装置，２は電子化された非定型の文書群，３は表示装置，４はキーボードやマウス等の入力装置，１１は関連日付抽出部，１２は頻度集計部，１３は頻度データ，１４は文書頻度表示制御部，１５は正規化処理部，１６は強調度表示処理部を表す。

【００１４】関連日付抽出部１１は，文書に関する日付情報を文書内容からパターンマッチにより自動的に抽出する処理手段である。頻度集計部１２は，日付ごとに関連する文書数を集計して頻度データ１３を得る処理手段である。文書頻度表示制御部１４は，日付ごとの文書頻度を，表示装置３などの出力手段に応じて強調して表示する制御手段である。

【００１５】正規化処理部１５は，文書頻度に応じて頻度情報の強調表示を行うために，年，月，曜日，日といった時間単位における文書頻度の分布に基づき，各文書頻度を正規化して強調度に変換する処理を行う。

【００１６】強調度表示処理部１６は，正規化処理部１５で得た強調度を用いて，出力装置に応じて文書頻度を強調して表示する。強調して表示する方法として，例えば頻度に応じて色を変える，明度または濃度を変える，頻度を表示する記号（文字を含む）や図形の大きさを変えるなどがある。

【００１７】文書の頻度を集計する時間単位は，入力装置４からの入力により，年，月，曜日または日といった指定が可能であり，頻度を表示する時間単位は随時切り換えることができる。

【００１８】本発明は以下のように作用する。まず，関

連日付抽出部１１は，電子化された文書群２の文書内容に記述されている日付表現に着目して，これらの日付表現を文字列のパターンマッチにより自動的に抽出し日付情報とし，頻度集計部１２は，日付情報ごとに関連する文書数を集計し，文書頻度表示制御部１４は，年／月／曜日／日単位といった複数の時間的観点から，各時間単位における文書数の分布を，表示装置３等の出力装置に応じて強調して提示する。

【００１９】以上の処理装置１における各処理部は，処理装置１のＣＰＵが実行するプログラムによって実現され，そのプログラムは適当な記憶媒体に格納して提供することができる。

【００２０】

【発明の実施の形態】以下，本発明の実施の形態の一例を説明する。

［関連日付抽出部］関連日付抽出部１１は，文書内容に記述された日本語および英語等の日付表現の文字列をパターンマッチにより，日付情報を抽出する。

【００２１】文書内容から関連する日付情報を抽出する処理は，１）数字表現の統一，２）パターンマッチ，３）曜日判定ルーチンの手順で行う。

〔１〕数字表現の統一

文書中には，日付の数字は種々の形式で記述される。そこで，パターンマッチによる日付情報の抽出を効率的に行うためには，パターンマッチの前に数字表現を統一しておくことが望ましい。数字表現の統一は，以下の手順で行う。

【００２２】１．全角数字０，１，…，９を，半角数字0,1,…,9に置き換える。

２．漢数字〇，一，…，九を，半角数字 0,1,…,9に置き換える。

３．“（数字１）十（数字２）”というすべての表現を，“（数字１）（数字２）”に置き換える。

【００２３】４．“（数字）十”という全ての表現を，“（数字）０”に置き換える。

５．“十（数字）”という全ての表現を，“１（数字）”に置き換える。

６．すべての“十”を“10”に置き換える。

【００２４】これにより，日付の表現に含まれる２桁までの数字は，全て半角数字の表現に変換される。

〔２〕パターンマッチ

パターンマッチの処理は，英語および日本語における日付を表す表現パターンを各文書内容に順次当てはめ，該当する日付を全て取り出す処理である。年が省略されている場合には，直前の日付表現パターンから得た年とし，それがない場合には現在（処理時点）の年とする。

【００２５】図２～図４は，パターンマッチにおいて用いる日付表現のパターンの例を示す。図２は英文における日付表現パターン例であり，図２（Ａ）は「日月年」の順序のパターン例，図２（Ｂ）は「月日年」の順序のパターン例，図２（Ｃ）は，期間を表す表現のパターン例である。

【００２６】図３は日本語文における日付表現パターン例であり，図３（Ａ）は数字のバリエーションの例，図３（Ｂ）は，年号（明治，大正，昭和，平成およびそれらの省略形）を使用するパターン例，図３（Ｃ）は西暦を使用するパターン例，図３（Ｄ）は，期間を表す表現例，図３（Ｅ）は慣用的表現の例である。

【００２７】図４はその他の日付表現パターン例を示す。図４に示す日付表現パターン例においては，パターンＡ「年－月－日」，パターンＢ「日－月－年」，またはパターンＣ「月－日－年」で記述されている可能性がある。このため，２０００年以前では，年数は「31」より大きいので，パターンＡとパターンＢまたはＣとの区別がつくが，パターンＢとパターンＣとの区別がつかない場合がある。その場合には，用途に応じて対象とする日付期間（例えば現在の前後１０年以内など）に属するパターンを全て抽出する。

【００２８】〔３〕曜日判定ルーチン

曜日判定ルーチンは，パターンマッチで抽出されたＸ年Ｙ月Ｚ日の曜日を計算する。

【００２９】曜日の計算は，最近何十年かのカレンダー等の表を用いて行ってもよいが，本実施の形態では，以下のような方法によって任意の年月日の曜日計算を行う。ｙ年ｍ月ｄ日に対して，

ＹＰ＝(y+2+int((y-1)/4)+int((y-1)/400))mod7

を計算する。ここで，int(x)はｘを越えない最大の整数，mod7は７で割った剰余を表す。また，

$\{M_1, M_2, \cdots, M_{12}\} = \{0,3,3,6,1,4,6,2,5,0,3,5\}$

とする。

【００３０】次の計算式で得られるｄｏｗが，ｙ年ｍ月ｄ日の曜日を表す。ただし，０：日曜，１：月曜，…，６：土曜である。

１．ｙ年が閏年（４で割り切れ，４００で割り切れない数）であって，ｍ＞２の場合：

ｄｏｗ＝（ＹＰ＋M_m ＋ｄ＋１）ｍｏｄ７

２．それ以外の場合：

ｄｏｗ＝（ＹＰ＋M_m ＋ｄ）ｍｏｄ７

なお，ここでは，非定型文書の文書内の日付表現から日付情報を抽出する場合を説明したが，文書に付加された情報の日付情報についても，同様に扱うことができる。

【００３１】以上，説明した日付情報の抽出処理は，「文書共有整理システム，共有文書管理装置および文書アクセス装置」（特願平８－２８１９４０号）において開示した関連日付抽出ルーチンに曜日判定ルーチンを加えて改良したものである。

【００３２】図５は，関連日付抽出部の処理フローチャートである。ステップＳ１０では，日付の集合Ｓを空集合とする。ステップＳ１１では，文書中に現れる数字表

現を統一する。

【００３３】ステップＳ１２では，文書中の各行について日本語，英語，その他の日付表現のパターン例を用いてパターンマッチを行う。ステップＳ１３では，文書中のその行に日付表現のパターンが存在するかどうかを判定する。パターンが存在する場合にはステップＳ１４の処理へ進み，パターンが存在しない場合にはステップＳ１５の処理へ進む。

【００３４】ステップＳ１４では，集合Ｓに検出した日付を追加する。ただし，同じ文書内の同じ日付がすでに集合Ｓに追加されている場合には追加せず，すでに追加されたものと異なる日付であれば，同一文書内にいくつあっても追加する。

【００３５】ステップＳ１５では，文書の終わりかどうかを判定する。文書の終わりでなければステップＳ１６の処理へ進み，文書の終わりであればステップＳ１７の処理へ進む。

【００３６】ステップＳ１６では，文書の次の行へ進み，ステップＳ１２以下の処理を文書が終わるまで繰り返す。ステップＳ１７では，検出した日付の曜日を判定し，集合Ｓに曜日情報を追加する。

【００３７】［頻度集計部］頻度集計部１２は，関連日付抽出部１１で抽出した全ての日付に対して，該当文書数を集計し，頻度データ１３を作成する。Ｘ年の文書数をｙｆ（Ｘ），Ｘ年Ｙ月の文書数をｍｆ（Ｘ，Ｙ），Ｕ曜日の文書数をｗｆ（Ｕ），Ｘ年Ｙ月Ｚ日の文書数をｄｆ（Ｘ，Ｙ，Ｚ）で表すとする。

【００３８】Ｘ年Ｙ月Ｚ日Ｕ曜日のデータに対しては，次のように文書数を集計する。

１．年単位：ｙｆ（Ｘ）＋＋

２．月単位：ｍｆ（Ｘ，Ｙ）＋＋

３．曜日単位：ｗｆ（Ｕ）＋＋

４．日単位：ｄｆ（Ｘ，Ｙ，Ｚ）＋＋

（＋＋は，文書数を１増やすことを表す。）

また，日付情報が期間を示すものである場合には，期間の最初と最後の日付の両方に対して集計する。さらに，$Max^{D}{}_{yf}$で時間的区間Ｄにおけるｙｆ（Ｘ）の最大値を，$Min^{D}{}_{yf}$で０より大きいｙｆ（Ｘ）の最小値を表すなどする。なお，後述する変換関数の種類によっては，$Min^{D}{}_{yf}$に０を含めてもよい。

【００３９】［文書頻度表示制御部］文書頻度表示制御部１４は，利用者の要求に応じて，例えば年，月，曜日，日の４つの単位で，文書頻度を表示画面上に強調して表示する。文書頻度の分布を容易に把握できるようにするため，表示装置に対応させて，頻度を色または濃度等に変えて強調度表示を行う。

【００４０】例えばカラーＣＲＴに表示する場合には，頻度を暖色から寒色への色相の変化として表示したり，モノクロの表示装置に表示する場合には，頻度を濃度（明度）の変化として表示したりする。また，色や濃度以外にも，表示する文字のフォントやサイズを頻度に応じて変えて表示するようにしてもよい。利用者からの要求があれば，表示単位の年／月／曜日／日を切り換えて表示する。

【００４１】文書頻度表示の制御は，１）頻度－強調度変換（正規化），２）強調度表示の手順で行う。

〔１〕頻度－強調度変換（正規化）

正規化処理部１５は，利用者が指定した時間的区間における文書頻度の最大値と最小値から，頻度－強調度の正規化関数を作成する。この正規化関数を用いて，各頻度を強調度に変換する。ここで，強調度とは［０，１］の実数である。

【００４２】具体的には，頻度－強調度の変換を以下のように行う。利用者が対象としたい時間区間をＤとして，非負整数による頻度範囲（例えば［$Min^{D}{}_{yf}$，$Max^{D}{}_{yf}$］の範囲）を，強調度［０，１］の実数に正規化を行う。変換関数は，頻度ｆ，頻度の最小値ｍｉｎ（＝$Min^{D}{}_{yf}$），頻度の最大値ｍａｘ（＝$Max^{D}{}_{yf}$）の関数Ｃｏｎｖ（ｆ，ｍｉｎ，ｍａｘ）として表すことができる。Ｃｏｎｖ（ｆ，ｍｉｎ，ｍａｘ）は，ｆに関して単調増加する関数であれば，用途や変換方法に応じていろいろと選ぶことができる。すなわち，ｆ１≦ｆ２であれば，Ｃｏｎｖ（ｆ１，ｍｉｎ，ｍａｘ）≦Ｃｏｎｖ（ｆ２，ｍｉｎ，ｍａｘ）となるような関数である。例えば，次のような関数が考えられる。

【００４３】１．線型変換

Ｃｏｎｖ（ｆ，ｍｉｎ，ｍａｘ）＝（ｆ－ｍｉｎ）／（ｍａｘ－ｍｉｎ）

この線型変換は，頻度にばらつきがない場合に適する。

【００４４】２．対数変換

Ｃｏｎｖ（ｆ，ｍｉｎ，ｆａｘ）＝ｌｏｇ（ｆ－ｍｉｎ＋１）／ｌｏｇ（ｍａｘ－ｍｉｎ＋１）

対数変換は，ｍｉｎとｍａｘが桁違いの場合など，ｍｉｎとｍａｘとが非常に離れていて，特にｍａｘ側でばらつきがあるような場合に適する。

【００４５】このような変換関数を，システムが頻度の分布に応じて自動的に選択するようにしてもよく，また利用者にメニューで選択させるようにしてもよい。

〔２〕強調度表示

強調度表示処理部１６は，強調度に従って文書頻度を表示する。すなわち，［０，１］の実数による強調度を，出力装置に応じて色や濃度を変えることによって表示する。出力装置として２４ビットのカラーＣＲＴを想定すると，次のような強調度表示が考えられる。

【００４６】１．色による強調度表示

色による強調度表示では，例えば「（寒色）青色－緑色－黄色－赤色（暖色）」という色の連続的な変化と強調度とを対応させる。ＲＧＢについてそれぞれ８ビット（２５６が最大値）で色を表すとすると，強調度ｅに対して各色の強さは，図６に示すような変換関数で表すこ

とができる。

【0047】図7は，図6に示す変換関数による強調度と色との変換の対応関係を示す図である。図6に示す関数では，図7に示すように，強調度が0で青，強調度が0から0．5の間は青と緑の混合，強調度が0．5で緑，強調度が0．5から1．0の間は緑と赤の混合（黄），強調度が1．0で赤となる。

【0048】2．濃度による強調度表示強調度を色の変化ではなく，濃度の変化とする場合，例えば次のように濃度の変化と強調度とを対応させる。

【0049】red（e）＝blue（e）＝green（e）＝256（1－e）

このような関数で，白色（RGBいずれも256）から黒色（RGBいずれも0）への対応を与えることができる。

【0050】〔3〕表示単位の切り換え

表示単位として，年／月／曜日／日の単位を切り換えて表示を行う。

1．年単位の表示

利用者が指定したある年の集合$X \in \{x_1, x_2, \cdots, x_n\}$に対して，Xとyf（X）との関係を表示する。通常は，yf（X）が定義されている期間のXを対象とすることになる。

【0051】図8に，年単位の表示例を示す。一次元的に年と強調度を表示する。強調度を色または濃度を変化させて表示することにより，年単位での文書数の分布の推移を把握することができる。図8では，文書頻度の高い年ほど高濃度で表示するようにしており，これにより，ある文書群の年度別の文書頻度（数）について，93年をピークとする文書数の推移が容易に把握できる。

【0052】2．月単位の表示

利用者が指定したある（年，月）ペアの集合$(Y, M) \in \{(y_1, m_1), (y_2, m_2), \cdots, (y_n, m_n)\}$に対して，Y，Mとmf（Y，M）との関係を表示する。通常は，mfが定義されている期間の（Y，M）を対象とすることになる。

【0053】図9に，月単位の表示例を示す。二次元で年と月とを表示し，強調度を色や濃度の変化で表示する。これにより，月単位の分布や年単位の分布を容易に把握することができる。図9からは，例えば「対象文書の頻度は，毎年，夏になると増加する傾向がある。」ことが把握できる。

【0054】3．曜日単位の表示

図10に示すように，利用者が指定した時間的区間の文書頻度に対して，一次元的に曜日と強調度を表示する。これにより，例えば「ある文書の頻度は週の半ばに多い。」等，曜日ごとの文書数の分布を容易に把握することができる。

【0055】4．日単位の表示

図11に示すように，利用者が指定したY年M月のカレンダーの中に，その月の日単位の文書頻度を表示する。これにより，例えば「ある文書の頻度は週の半ばに多い。」とか，「ある文書の頻度は月末に多い。」等，月の中での文書数の分布を容易に把握することができる。

【0056】

【実施例】本発明は，テキストファイル等，一般の非定型文書に対して適用できるため，例えばオペレーティング・システム（OS）のファイルシステムに組み込むなどの直接的な応用が考えられる。

【0057】ここでは，検索システム，グループウェアといった既存の情報処理システムに組み合わせた場合の実施例を説明する。

［1］検索システムにおける検索結果の可視化システム

インターネットのホームページのような非定型データの検索手法として，全文検索がある。これは，データベースや特定のキーワードの検索とは異なり，文書内の任意の文字列に関して検索を行うものである。ただし，ある文字列が一致するからといって，必ずしもそれが求める文書とは限らない場合があるため，検索結果にはノイズが含まれる。また，検索対象となる文書の全体量が増大していること，さらに検索漏れをなくすために類似語をORで展開して検索すること等が要因となって，検索結果の文書数も膨大な量になる。このような，膨大でノイズの多い検索結果の中から有用な情報を探し出すために，検索結果の可視化を行うことが望まれる。

【0058】本発明は，このような検索結果の可視化に適用できる。具体的には，全文検索の検索結果である非定型文書群に対して本発明を適用し，文書頻度の強調度表示により検索結果の可視化を行う。図12は，全文検索システムの検索結果の可視化システムとして本発明を適用した場合の構成例を示す図である。

【0059】図12に示すデータベース検索エンジン21により，1990年から1994年までのある新聞の全記事（データベース22）を対象に検索する場合を想定する。

【0060】利用者からの検索要求により，データベース検索エンジン21が，データベース22を全文検索し，該当する検索結果を本発明の文書表示装置に与える。関連日付抽出部11は，その検索結果（文書群）を得て，文書内に記述されている日付表現からパターンマッチにより日付情報を抽出し，頻度集計部12により，日付情報ごとに関連する文書数を集計し，頻度データ13とする。文書頻度表示制御部14は，年／月／曜日／日単位等の各単位における文書数の分布を，出力装置に応じて強調して表示し，利用者にビューとして提示する。

【0061】ここでは，月単位のビューおよび日単位のビューを出力するものとする。検索対象が新聞記事なので文書頻度には極端なばらつきがないため，正規化変換には線型変換を使っている。強調度の表示には色の変化

（寒暖）を用いている。

【００６２】利用者は，月単位のビューにおいて月のセルを，例えばマウス等によりクリックして選択すると，該当する月の日単位のビューが表示される。これによって，月単位のビューで全体の傾向を把握し，さらに日単位のビューにより，より詳しい傾向を把握し，さらには検索結果そのものへアクセスすることができる。このように，本発明では，利用者の指示入力により，ビューの単位を変えて検索結果を表示することから，利用者は，複数の時間的観点から検索結果を容易に把握することができる。

【００６３】図１３および図１４は本実施例におけるビューの例を示す図である。図１３は，「食中毒」という文字列を含む記事を検索した結果を月単位のビューで表示した例を示している。このビューにより，「食中毒」に関する記事は，傾向として夏に多いということがわかる。検索結果をさらに絞り込むために条件を追加する場合，通常のデータベース検索の場合と同様に，データベース検索エンジン２１に絞り込みの条件を送る。

【００６４】図１４は，「高速道路」および「渋滞」を含む記事を検索した結果を１９９０年の４月の日単位のビューで表示した例を示している。このビューにより，高速道路の渋滞に関する記事は，傾向として週末／休日前に多いことがわかる。

【００６５】［２］文書共有管理システム（グループウェア）における共有文書の可視化システム

グループウェアの目的の一つは，各利用者の文書情報を共有することにあるが，本発明は，このようなグループウェアの文書情報の可視化に適用できる。

【００６６】具体的には，「文書共有整理システム，共有文書管理装置および文書アクセス装置」（特願平８－２８１９４０号）の文書共有整理システムと組み合わせ，文書共有整理システムの文書蓄積および再利用の拡張をするものである。

【００６７】まず，「文書共有整理システム，共有文書管理装置および文書アクセス装置」（特願平８－２８１９４０号）を説明する。この文書共有整理システムは，例えばインターネット（イントラネット）等のネットワークを利用して文書を共有するシステムであって，利用者が有用と思われる情報を簡単な操作で付加情報とともにグループの共有マシンに登録することができ，それらの文書群の情報を整理して提示し，更新することを可能にしているものである。

【００６８】図１５は，文書共有整理システムの共有文書の可視化システムとして本発明を適用した場合の構成例を示す図である。ローカルマシン４０では，入力装置４１により入力・編集された文書を所定の文書構造に変換し，送信手段４２を介してネットワークを通じて共有マシン５０に送る。

【００６９】共有マシン５０では，受信手段５１により，受信した文書に文書ＩＤを付与して共有文書群５２に蓄積し，文書ＩＤ等をインデックス登録手段５３へ送る。インデックス登録手段５３では，文書ＩＤ等をインデックスデータ５４に登録する。ビュー生成手段５５では，文書群を時間または利用者の文書へのアクセスに応じて整理し，共有文書群５２に蓄積された文書群またはそのインデックスデータを時間順や利用者のアクセスに応じて自動的に整理した表示出力（ビュー）を表示装置４３に提示する。

【００７０】本実施例は，インデックスデータ５４に，文書に関連する日付ごとの文書頻度を示す頻度データを追加し，ビュー生成手段５５において，共有文書群５２に対するビュー（可視化）の一つとして，年／月／曜日／日といった時間単位ごとの文書の頻度情報を表示することができるようにしたものである。

【００７１】このため，インデックス登録手段５３に，本発明に係る関連日付抽出部１１および頻度集計部１２を組み込み，インデックスデータ５４の一つとして頻度データ１３を管理する。ビュー生成手段５５では，ローカルマシン４０からの文書頻度の要求により，文書頻度表示制御部１４によって，年／月／曜日／日といった時間単位ごとの文書頻度情報を表示する。

【００７２】このように，グループのメンバー（利用者）が登録した文書群を共有するときのビューの一つとして本発明を適用することによって，共有文書群５２のさらに容易な把握が可能になる。

【００７３】

【発明の効果】以上説明したように，本発明は以下のような効果を奏する。

１．テキストファイルなどの非定型な文書群は身の回りにあふれている。それらの分布を時間に沿って可視化することができる。

【００７４】２．日付ごとの文書の頻度は濃度や色の変化により可視化されるため，時間的傾向や変化の推移を一目で把握することができる。特に，検索システムと組み合わせることで，あるトピックに関する時間的推移を把握することができる。

【００７５】３．文書に関連する日付の自動抽出は，日本語または英語の文書に対応している。日頃アクセスする大半の文書に対して適用が可能であり，言語に固有な日付表現のパターンを追加することで，種々の言語にも応用が可能である。

【００７６】４．テキストファイル群だけでなく，全文検索の検索結果や，グループウェアの共通文書群に対しても適用が可能である。

５．文書頻度表示処理に関しては，定型データに対しても適用できる。

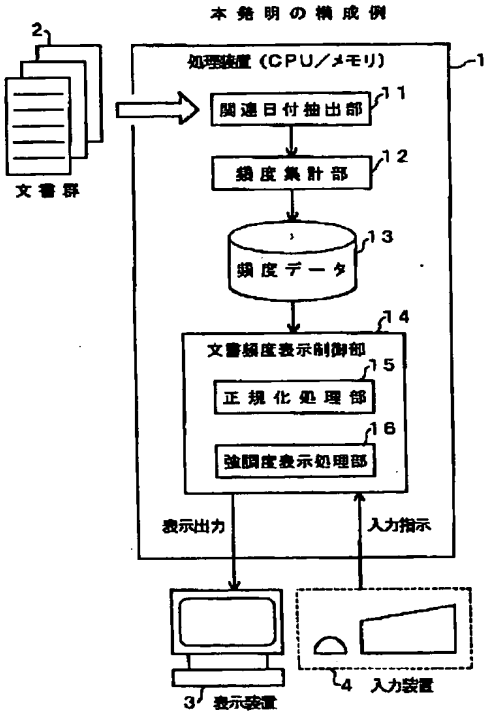
【図面の簡単な説明】

【図１】本発明の構成例を示す図である。

【図２】英文における日付表現パターン例を示す図であ

る。
【図３】日本語文における日付表現パターン例を示す図である。
【図４】その他の日付表現パターン例を示す図である。
【図５】関連日付抽出部の処理フローチャートである。
【図６】強調度を色に変換する変換関数の例を示す図である。
【図７】強調度と色の変化との変換の対応関係を示す図である。
【図８】年単位の表示例を示す図である。
【図９】月単位の表示例を示す図である。
【図１０】曜日単位の表示例を示す図である。
【図１１】日単位の表示例を示す図である。
【図１２】本発明を全文検索システムの検索結果の可視化システムとして本発明を適用した場合の構成例を示す図である。
【図１３】ビューの例を示す図である。
【図１４】ビューの例を示す図である。
【図１５】文書共有整理システムの共有文書の可視化システムとして本発明を適用した場合の構成例を示す図である。
【符号の説明】
１　処理装置
２　文書群
３　表示装置
４　入力装置
１１　関連日付抽出部
１２　頻度集計部
１３　頻度データ
１４　文書頻度表示制御部
１５　正規化処理部
１６　強調度表示処理部

【図１】



【図２】

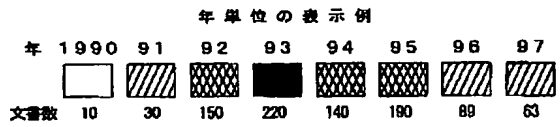
英文における日付表現パターン例

(A)
```
23 March 1998       (基本)
24 Mar. 1996        (月名が省略系)
29 March 96         (年が下２桁)
23 March '96         (年が ' および下２桁)
15th May 1996       (日に st,nd,rd,th がつく)
17th of May in 1886  (of,in のような前置詞がはさまっている)
Thu, 11 May 1995    (曜日が付加されている：ここでは無視する)
23 March        (年の省略)
```

(B)
```
September 2, 1996     (基本)
Sep. 3, 1996          (月名が省略系)
September 2, 96     (年が下２桁)
September 2, '96     (年が ' および下２桁)
May 16th, 1996       (日に st,nd,rd,th がつく)
September 2 in 1996     (in のような前置詞がはさまっている)
September 2     (年の省略)
```

(C)
```
22-25 September 1997   (基本)
January 11-12, 1997    (基本)
January 13 — 14, 1997 (デリミタが — )
Jan. 28 - Feb. 10, 1997 (月にまたがる場合)
Dec. 27, 1996  - Jan. 14, 1997 (年をまたがる場合)
```

【図８】



【図１０】

【図3】

日本語文における日付表現パターン例

(A)

10：十、一〇
12：十二、一二
20：二十、二〇
23：二三、二十三

(B)

平成8年6月25日　（基本）
平8年6月25日　（年号が省略形）
1996(平成8)年10月17日　（西暦との組み合わせ）
H7.8.4　（年号が省略形、デリミタが.）
H8-4-5　（年号が省略形、デリミタが-）
H8/4/6　（年号が省略形、デリミタが/）

(C)

1996年10月17日　（基本）
1996年10月17日(木)　（曜日が入っている：無視する）
96年10月17日　（年が下2桁）
6月29日　（年の省略）

(D)

平成8年10月2～10日
平成8年10月2日～10日
平成8年10月25日～11月10日
平成8年12月25日～平成9年1月10日
1995(平成7)年10月31日(木)～11月1日(金)

(E)

平成8年　2月末　（月の最終日）
平成8年　2月中旬　（15日とみなす。他に初旬=5日，下旬=25日）

【図4】

その他の日付表現パターン例

96/7/1　（1996年7月1日）
1996/7/3　（1996年7月3日）
96/7/6 - 8　（1996年7月6日から1996年7月8日）
96/7/5 - 8/13　（1996年7月5日から1996年8月13日）
96.6.3　（1996年6月3日）
8/18　（8月18日、年は直前のパターンの年または現在の年）
8/5--9　（8月5日から8月9日、年は直前のパターンの年または現在の年）
8/5--9/14　（8月5日から9月14日、年は直前のパターンの年または現在の年）
9/5　（3月5日または5月3日）
1996-8-6　（1996年8月6日）
960821　（1996年8月21日）
19960822　（1996年8月22日）

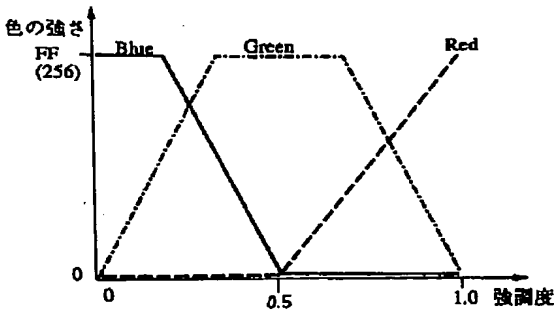
【図6】

$$red(e) = \begin{cases} 0, & \text{for } e =< 0.5 \\ 256(e-0.5)/0.5, & \text{for } 0.5 < e \end{cases}$$

$$green(e) = \begin{cases} 256e/0.4, & \text{for } e =< 0.4 \\ 256, & \text{for } 0.4 < e < 0.75 \\ 256(1-e)/0.25, & \text{for } 0.75 =< e \end{cases}$$

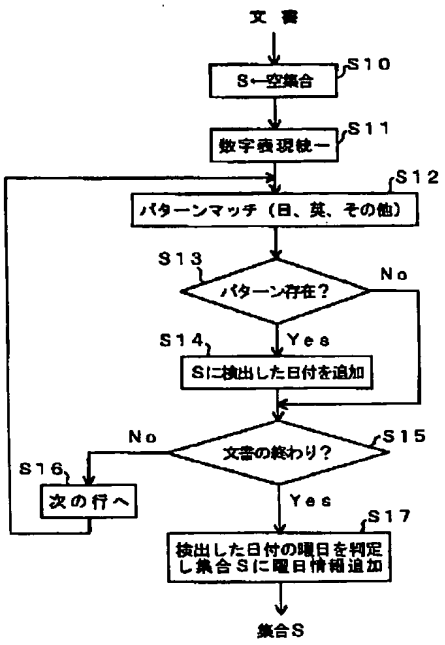
$$blue(e) = \begin{cases} 256, & \text{for } e =< 0.25 \\ 256(0.5-e)/0.25, & \text{for } 0.25 < e < 0.5 \\ 0, & \text{for } 0.5 =< e \end{cases}$$

【図7】



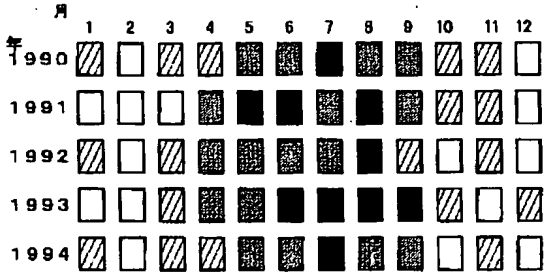
【図5】

関連日付抽出部の処理フローチャート

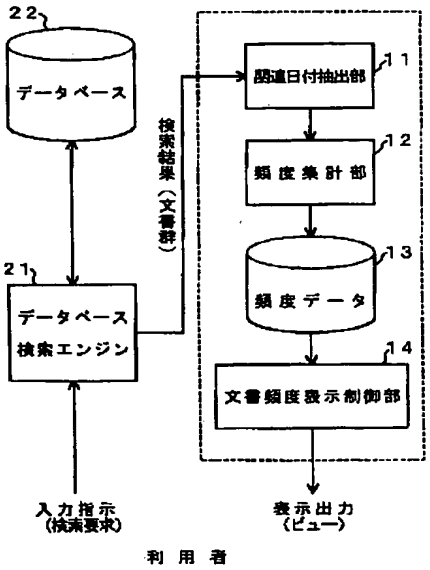


【図9】

月単位の表示例

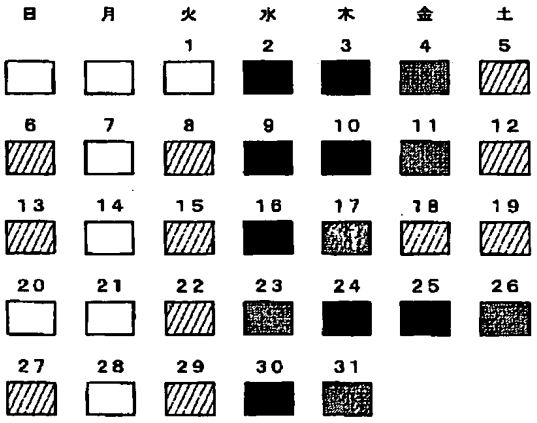


【図12】

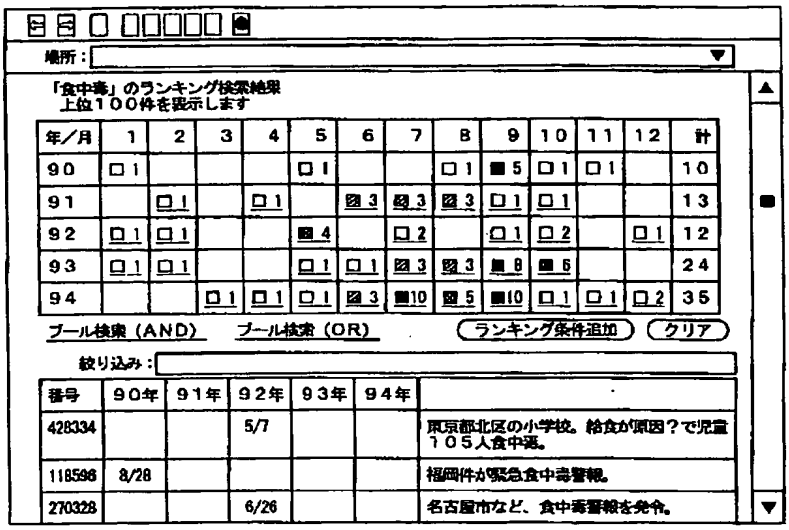


【図11】

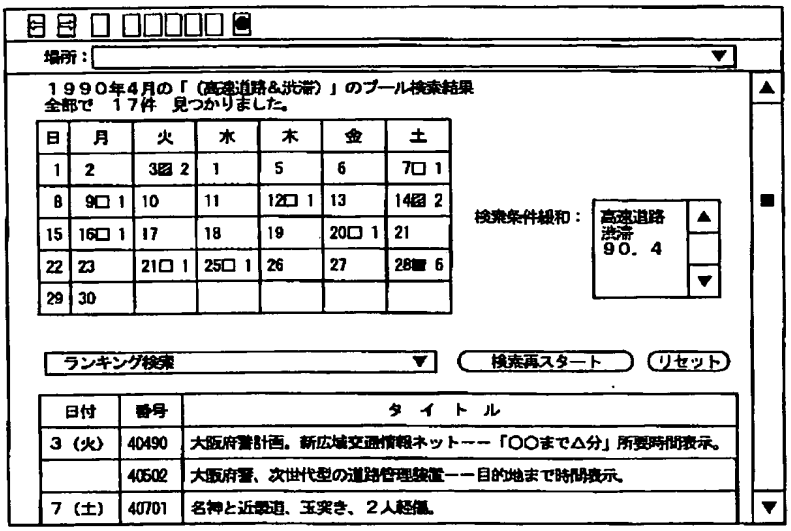
日単位の表示例



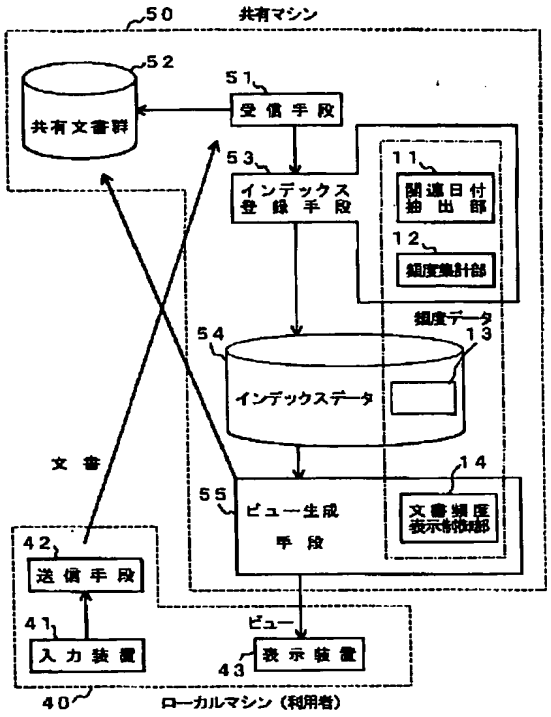
【図１３】


場所：
「食中毒」のランキング検索結果
上位１００件を表示します
年／月 1 2 3 4 5 6 7 8 9 10 11 12 計
90 1 5 1 1 10
91 1 3 3 3 1 1 13
92 1 1 4 2 1 2 1 12
93 1 1 1 1 3 3 8 6 24
94 1 1 1 3 10 5 10 1 1 2 35
ブール検索（AND）
ブール検索（OR）
ランキング条件追加
クリア
絞り込み：
番号 90年 91年 92年 93年 94年
428334 5/7 東京都北区の小学校。給食が原因？で児童１０５人食中毒。
118596 8/28 福岡件が緊急食中毒警報。
270328 6/26 名古屋市など、食中毒警報を発令。

【図１４】


場所：
１９９０年４月の「（高速道路&渋滞）」のプール検索結果
全部で　１７件　見つかりました。
日 月 火 水 木 金 土
1 2 3 2 1 5 6 7 1
8 9 1 10 11 12 1 13 14 2
15 16 1 17 18 19 20 1 21
22 23 21 1 25 1 26 27 28 6
29 30
検索条件絞和：
高速道路
渋滞
９０．４
ランキング検索
検索再スタート
リセット
日付 番号 タ イ ト ル
3（火） 40490 大阪府警計画。新広域交通情報ネットーー「○○まで△分」所要時間表示。
40502 大阪府警、次世代型の道路管理装置ーー目的地まで時間表示。
7（土） 40701 名神と近畿道、玉突き、２人経傷。

【図15】


50
共有マシン
52
共有文書群
51
受信手段
53
インデックス登録手段
11
関連日付抽出部
12
頻度集計部
頻度データ
13
54
インデックスデータ
14
55
ビュー生成手段
文書頻度表示制御部
文書
42
送信手段
41
入力装置
ビュー
43
表示装置
40
ローカルマシン（利用者）